DOCUMENT RESUME

ED 266 638 FL 015 452

AUTHOR Raatz, Ulrich
TITLE Investigating Dimensionality of Language Tests--A New Solution to an Old Problem.
PUB DATE 85
NOTE 15p.; In: Practice and Problems in Language Testing 8. Papers presented at the International Language Testing Symposium of the Interuniversitare Sprachtestgruppe (IUS) (8th, Tampere, Finland, November 17-18, 1984); see FL 015 442.
PUB TYPE Reports - Research/Technical (143) -- Speeches/Conference Papers (150)

EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS Comparative Analysis; Data Interpretation; *Evaluation Methods; *Factor Analysis; Foreign Countries; *Language Tests; *Latent Trait Theory; Models; Second Language Learning; *Test Validity
IDENTIFIERS West Germany

ABSTRACT

The use of factor analysis in examining the dimensionality of language tests is questioned and found to be problematic because it has a number of weak points which make it possible to manipulate results. Some of these weaknesses occur in the examinee selection sample, test selection, and the methods of extraction and rotation chosen. The use of the Classical Latent Additive Test Model, a latent-trait model comparable to the Rasch Model, is considered prefereble because of its more empirical basis. In application to three language tests--the C-Test, the oral interview, and the Delta-Test (Duisberg English Language Test for Advanced Students) battery of 21 subtests--at a German university, the Classical Latent Additive Test Model was found to be effective in assessing test dimensionality and in determining the optimum combination of Delta-Test subtests for a variety of assessment purposes. A number of tables showing the results are included. (MSE)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

BEST COPY AVAILABLE

ED266638

INVESTIGATING DIMENSIONALITY OF LANGUAGE TESTS - A NEW SOLUTION TO AN OLD PROBLEM

Ulrich Raatz (University of Duisburg)

1. Introduction

The content of this paper is factorial validity of language tests. Two of the most interesting questions in this area are the questions about the dimensionality of language and about the dimensionality of language tests.

These two questions and the methods of language teaching are closely connected. A given structure demands specific measurement techniques and teaching procedures. Thus a structuralist model of language with the components listening, speaking, reading and writing will produce listening comprehension tests, oral tests, reading comprehension tests and writing tests. An integrated approach demands global (holistic) language tests such as the interview, the essay or the cloze test.

At the present time these two models of language are in competition with each other. Each is based on a theory of language, each is confirmed by empirical data, each has led to the development of specific testing procedures.

Which model is correct? Or, perhaps, to put it more accurately, which model is a better description of reality?

Perhaps it is possible to solve the problem on the basis of theoretical considerations alone. Another approach, however, would be to use empirical methods.

Such an empirical examination of the proposed hypotheses about the structure of language tests, and consequently also about the structure of language itself, is generally conducted by administering a number of different language tests to a sample of subjects and performing factor analysis on the test results. The factors discovered enable conclusions to be drawn in favour of one of the above-mentioned models.

This type of research design can be illustrated by a number of studies. The results of these studies cannot be said to be unanimous. They differ in the existance of a general factor, in the number of common factors, and in the interpretation of these. In short, they do not help us to decide which of the hypotheses about the structure of language is the "correct" one. How can this lack of unanimity be explained?

Detailed examination of the studies reported above shows considerable differences in the methods and the evaluation techniques

FL015452

U.S. DEPARTMENT OF EDUCATION
NATIONAL INSTITUTE OF EDUCATION
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

X This document has been reproduced as received from the person or organization originating it.

[ ] Minor changes have been made to improve reproduction quality.

• Points of view or opinions stated in this document do not necessarily represent official NIE position or policy.

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY

J. Tommola

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

used. Some of the samples are homogenous, some are heterogenous. Each study uses different tests. But the main problem is the use of factor analysis. This statistical method is problematical for solving the problem of dimensionality. Why?

2. A brief overview of factor analysis

Factor analysis is a method belonging to the area of descriptive statistics. It permits the researcher to reduce large quantities of information (for instance the results derived from a large number of tests) to certain common underlying aspects or dimensions; the so-called factors.

A factor analytic study using the methods originally developed by Thurstone (1947) normally proceeds in the following steps:

1. A number of tests are selected which are to be described more parsimoniously, namely by reduction to a small number of underlying factors.

2. A sample of examinees is selected.

3. The tests are administered to this sample.

4. All intercorrelations between the tests are calculated.

5. The individual factors are "extracted", in other words the relevant loadings are determined for one factor after another. For this problem there are a number of different methods available.

6. The number of factors finally extracted is determined by a so-called cut-off criterion (e.g. Eigenvalue-criterion).

7. Rotation is performed. This is done because factor analysis has an infinite number of equally satisfactory solutions which can all be converted into each other by rotation. Every single solution is correct.

   During rotation the number of factors generally remains constant; what changes is the amount of variance explained by the individual factors. So a g-factor can turn into a group factor.

   The criteria used to determine which of the infinite number of possible rotations is the "correct" one can be either psychological (intuitive) or analytic. A researcher using psychological criteria selects the solution which best supports his original hypothesis. Analytic rotations are determined by mathematical methods. The best-known of these is the varimax rotation to "simple structure".

BEST COPY AVAILABLE

8. Finally the factors are named. Generally the names chosen are determined by the tests which have the highest loadings on the individual factors.

3. Some problems with factor analysis

Factor analysis has a number of weak points so that it is possible to manipulate the results. The most important of these are:

1. the selection of the sample of examinees
2. the selection of the tests
3. the method of extraction chosen
4. the method of rotation chosen

ad 1: Kalverum (1965,1970) showed theoretically that the results obtained by a factor analysis could be dependent on the sample. If a sample is extremely heterogeneous one will tend to find only few factors, in extreme cases only a g-factor. If, on the other hand, the samples are homogeneous more factors will emerge.

ad 2: The reason for this is obvious. If, for instance, no oral test is included in the test battery, a factor analysis of the battery cannot yield an oral factor.

Only if a number of similar oral tests are added to the battery, and thus determine a large part of the variance in the matrix, is it probable that an independent oral factor will emerge. Obviously, then, the detection of certain factors is the result of the selection of tests entered into the factor analysis.

ad 3: The use of different methods of extraction of the factors can produce different factor patternings. This is true for the most commonly used common factor method, for the principal component method, and the method invented by Spearman.

ad 4: The rotation performed after extraction has taken place is a matter of personal choice. There are an infinite number of correct solutions. Thus, a "psychological" rotation can be performed, aimed at producing an optimal approximation to a particular theoretically-determined structure. Even if analytic rotation is used, there are a number of possibilities available, some of which tend more to a g-factor, some more to group factors.

Perhaps these weak points account for the different results of factor analytic studies of dimensionality of language tests. In fact, no researcher will manipulate the results, but everyone works in a different manner. You see, that factor analysis is not the best method for solving the problem of language dimensions.

BEST COPY AVAILABLE

4. The CLA-model

I propose a different solution. This is provided by the CLA-model proposed by Moosbrugger and Müller first in 1980 (1982). CLA stands for Classical Latent Additive Test Model, and, as the name implies, this is a latent-trait model, comparable to the Rasch-Model.

In its original form, this CLA-model is applicable to items which are least ordinally scaled. I have examined the use of this model for tests, in particular for investigating whether a number of tests are undimensional or not.

The CLA-model has the same characteristics as the Rasch-model:

- the true score is split into an ability parameter and a difficulty parameter;
- it has what is called specific objectivity; that means that the item difficulties and subject abilities can be estimated independently of each other;
- if the model fits, then the total score is undimensional and interval-scaled;
- model fit can be tested empirically.

I shall now present the basic characteristics of the model and illustrate its use with three examples.

The CLA-model has the following basic assumptions:

A test consists of c independent, equivalent parts (items, superitems, subtests). X_{pi} is the raw score of subject p (p = 1...N) in test part i (i = 1...c).

Just as in classical test theory we have the assumption that

$$X_{pi} = \tau_{pi} + \varepsilon_{pi} \qquad (1)$$

where τ is the true score and ε is the error component. The second assumption is

$$E(\varepsilon_{pi}) = 0 \qquad (2)$$

It follows, therefore, that

$$E(X_{pi}) = \tau_{pi} \qquad (3)$$

i.e the expected value for X_{pi} is the true score τ_{pi}.

In the CLA-model we split the true score into two components:

$$\tau_{pi} = \xi_p - \beta_i \qquad (4)$$

ξ_p is the ability of subject p; β_i is the difficulty level of item i.

BEST COPY AVAILABLE

Just as in the Rasch model we assume local stochastic independence, both on the superitem level (independence of content) and on the subject level (no copying).

From these assumptions we can now derive the "specific independence" of the model.

If we take two persons p and q and an arbitrary item i, then it follows for the difference of the two expected X-values:

$$E(X_{pi}) - E(X_{qi}) = \xi_p - \xi_q \qquad (5)$$

This difference is independent of the item.

If we take one arbitrary person p and two items i and j, then it follows for the difference of the two expected X-values:

$$E(X_{pi}) - E(X_{pj}) = \beta_j - \beta_i \qquad (6)$$

This difference is independent of the person.

The item characteristics in this model have the following form:

Figure 1

Item characteristics in the CLA-model

The model parameters ξ and β can be estimated as follows:

$$\xi_p' = \frac{1}{2} \sum_i (X_{pi} + \beta_i) \qquad (7)$$

BEST COPY AVAILABLE

or, if we normalise the β-values so that their sum equals 0, we can simplify the equation to:

$$\xi_p' = \frac{1}{c} \sum_i X_{pi} \qquad (8)$$

For the items we have:

$$\beta_i' = \frac{1}{N} \sum_p (\xi_p - X_{pi}) \qquad (9)$$

In this formula ξ_p has to be estimated by the values derived from (8).

In the case of a test battery with c subtests with n items each these formulae can be simplified. Let X_{pi} be the raw score of the subject p in subtest i, X_{ptot} the total score, $\bar{X}_i$ the mean score in subtest i and X_{tot} the overall mean. Then we find from (8) that:

$$\xi_p' = \frac{1}{c} X_{ptot} \qquad (10)$$

and from (9)

$$\beta_i' = \bar{X}_{tot} - \bar{X}_i \qquad (11)$$

There are a number of ways in which one can check for model fit with the CLA model:

1. The subject group is split into two parts using some arbitrary criterion. For each group the item parameters (β's) are calculated separately. The pairs of betas (one for each group) should show identical values which have only chance differences. For each pair the deviations are tested with the t-statistic; for the items overall with the chi-square statistic.

2. The item characteristics for all items should be parallel. This means, of course, that each superitem must have the same variance and item discrimination index. This condition is satisfied if TUKEY's test for non-additivity is non significant.

If these model checks turn out to be negative, the items or subtests which do not fit the model can be easily identified and

BEST COPY AVAILABLE

eliminated. If the model fits, the sum scores of the items, the superitems or the subtests are assumed to be undimensionally scaled on an interval scale.

5. 1st example: C-Test

In Duisburg a project is currently in progress which is supported by the "Arbeiterwohlfahrt", a social service agency. The aim of this project is to enable gifted Turkish girls to obtain a higher educational qualification. Around 30 girls attend courses for two years, and at the end of this time they take the "Mittlere Reife" examination. The number of applicants for this programme is much larger than the number of places available, and therefore selection procedures are necessary. Former school grades are the one selection criterion; the other is proficiency in the second language, German. This proficiency has, up to now, been assessed on the basis of a home-made test. The people in charge of this project asked us whether we could develop a German C-Test which would either provide additional information or even replace the present test procedure.

We constructed a C-Test with 5 parts. Each part had 20 deletions.

26 Turkish girls took the tests at the first test session in March 1984.

If we use the CLA-model on the C-Test data for test evaluation we begin by deriving the beta values (eq. 9) given in Table 1 as item difficulty indices for the C-Test parts.

Table 1

CLA-difficulties of the 5 superitems and the tests for goodness of fit

	β	β_{sup}	β_{inf}	t-test
C 1	-5,6	-5,0	-6,2	1,51 n.s.
C 2	-0,6	-0,7	-0,4	0,55 n.s.
C 3	2,9	2,7	3,2	0,98 n.s.
C 4	1,1	0,8	1,4	0,84 n.s.
C 5	2,1	2,2	1,9	0,45 n.s.

$\chi^2 = 4,45$ $(50\% > p > 25\%)$

Tukey: $F = 10,3$ $(p < 0,2\%)$

BEST COPY AVAILABLE

For the first test of model fit we separate the group on the basis of their C-Test scores into a higher and a lower half. Then the item parameters are determined for each group separately. These must be identical if the model fits.

The results are given in Table 1, together with the results of the t-tests. As one can see there is no pair for which the differences in the beta values are statistically significant, and the chi-square test over all values is not significant.

But the second model fit is unsatisfactory since the TUKEY test was statistically significant ($p < 0,2\%$). It is obviously superitem 1 which is causing problems, as inspection of the t-tests shows. We therefore decide that item 1 ought to be eliminated.

Table 2

CLA-difficulties of the 4 superitems and the tests for goodness of fit

	β	β_{sup}	β_{inf}	t
C 2	-2,0	-1,9	-2,0	0,20 n.s.
C 3	1,5	1,3	1,8	0,92 n.s.
C 4	-0,3	-0,3	-0,2	0,14 n.s.
C 5	0,7	0,9	0,4	0,84 n.s.

$\chi^2 = 1,61$ ($p > 50\%$)
Tukey: $F = 0,51$ ($p \approx 70\%$)

Table 2 shows the newly calculated item difficulties β, and the results of the two checks for model fit. Once again there are no significant individual differences, and the overall chi-square test is not significant. Now the TUKEY test is not statistically significant either. We can assume model fit. This means:

- the C-Test consisting of parts 2 to 5 is assumed to measure unidimensionally;
- the total scores on the C-Test are assumed to be scaled on an interval level.

6. 2nd example: Interview

In Duisburg, the oral performance of all students of Anglistik is measured at the end of the 3rd or 4th semester. This test is a mixture of prepared roleplay and interview in a group of 3 students with random contest and takes about 20 minutes (see Swith and

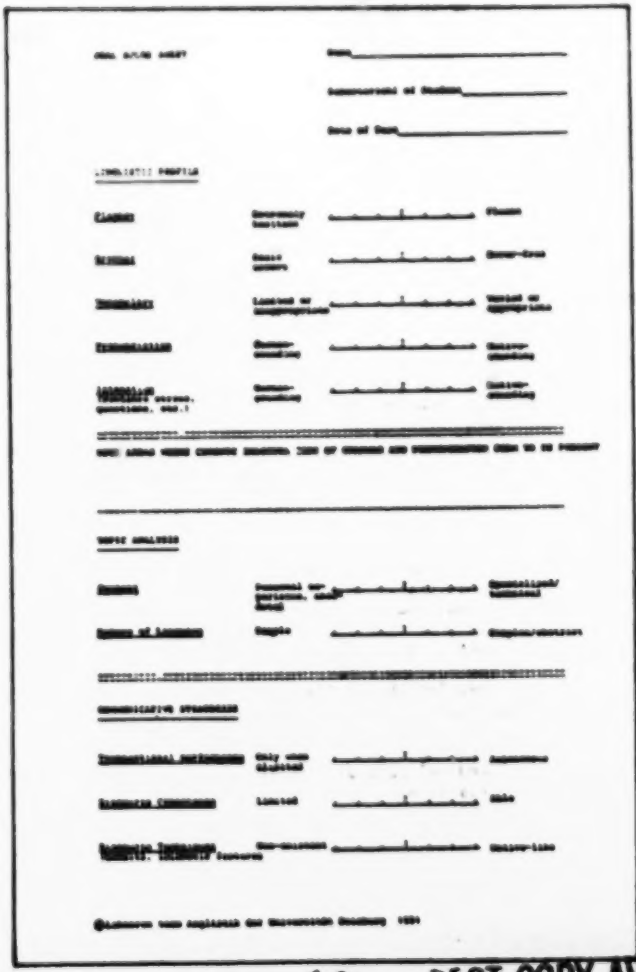
BEST COPY AVAILABLE

Klein-Braley, 1981, and Smith, 1982).

The randomly selected raters rate the characteristics of the oral competence of each student independently (on 6-point rating scales) using a score sheet with 10 categories.

Figure 2

Oral score sheet



BEST COPY AVAILABLE

In this Figure you see the 10 categories with the rating scales and the descriptions of the extreme points of these scales.

Our question was: how many dimensions does this language profile have. We performed the CLA-analysis for a group of 77 students taking the test in autumn 1982. For each student, two ratings had been given by the raters, so we calculated the mean for every scale.

The precondition was fulfilled: on each rating-scale the raw score ranged from 1 to 6 points.

The results of the CLA-analysis of the 10 scales are presented in Table 3.

Table 3

Results of CLA-analysis of oral test

	β_{sup}	β_{inf}	t-test
Fluency	-.61	-.58	.12 ns
Grammar	.03	-.19	.84 ns
Vocabulary	-.49	-.19	1.37 ns
Pronunciation	.22	-.38	3.06 ns
Intonation	.51	-.08	3.02 ns
Content	1.64	1.67	.06 ns
Nature of language	.69	.62	.30 ns
Transactional perform.	-1.19	-.74	1.63 ns
Discourse competence	-1.00	.69	1.46 ns
Discourse techniques	.20	.56	1.67 ns

$\chi^2 = 13.5$ ($P < 10\%$)
Tukey: $F = 0{,}62$ ($P \approx 43\%$)

The first column shows the names of the scales, the second and third the estimations of the difficulty indices, beta, in the upper and lower scoring group, and the 4th the result of the t-tests. Only two t-tests, for the scales "pronunciation" and "intonation", were significant at the 0,1% level. The overall χ^2-Test was nearly significant (10 %), the TUKEY-Test for additivity was non-significant (43%). These results showed that the scale of the 10 categories was not unidimensional.

Therefore we deleted the two categories "pronunciation" and "intonation" and calculated a second CLA analysis for the remaining 8 categories (Table 4).

BEST COPY AVAILABLE

Table 4

Results of CLA-analysis of oral test
with reduced categories

	β_{sup}	β_{inf}	t-test
Fluency	-.51	-.64	.47 ns
Grammar	.12	.25	1.34 ns
Vocabulary	-.39	-.25	.63 ns
Content	1.73	1.61	.21 ns
Nature of language	.78	.56	1.00 ns
Transactional perform.	-1.10	-.80	1.11 ns
Discourse competence	-.91	-.75	.76 ns
Discourse techniques	.29	.50	.98 ns

χ^2 = 6.5 (P > 50%)
Tukey: F = .03 (P ≈ 86%)

The t-tests, the overall χ^2-test and the test of additivity are not significant. This means that the scale consisting of the remaining 8 categories is assumed to be unidimensional. All these categories measure the same construct on an interval scale. The two categories "pronunciation" and "intonation" constitute perhaps one or two additional dimensions.

7. 3rd example: Delta-Test

In October 1984, we gave a modified form of the Delta-Test (Duisburg English Language Test for Advanced Students, Klein-Braley and Lück 1980) as a placement test to 65 beginners.

This test consisted of the following parts, each with 25 items.

BEST COPY AVAILABLE

Table 5

Modified Delta-Test
Duisburg, Oct 1984

Voc 1	slot-filling	active vocabulary
Voc 3a	MC 5 distractors	passive vocabulary
Voc 3b	MC 5 distractors	passive vocabulary
Gra 1	MC 5 distractors	variety of problem areas
Gra 2	slot-filling	verb tenses and aspects
Gra 3	slot-filling	prepositions and verb complem.
C 2		
C 3	5 texts	
C 4	slot-filling	integrative
C 5		
C 6		
Dict 1		
Dict 2	182 word text	
Dict 3	7 parts with	
Dict 4	25 words each	difficult formal prose
Dict 5	(beginning with the	integrative
Dict 6	8th word)	
Dict 7		
Trans 1	25 short sentences to be translated from G to E MC 4 distractors	formal prose integrative
Trans 2		
Trans 3	2 texts to be translated from G to E each with 25 parts right/wrong	formal prose integrative

The tests for grammar and vocabulary were discrete-point item tests with MC-or openended items. The holistic tests were a C-Test with 5 parts, a 183-word dictation and 3 translations, one of these in MC-format.

Our question was again: is this battery of 21 language tests unidimensional or not?

In the following Table 6 you see the results of some of the calculations we made.

BEST COPY AVAILABLE

Table 6

Results of CLA-analysis of the modified Delta-test

combination	number of tests	number of sign t-tests	result of χ^2-test	result of TUKEY-test
all tests	21	12	χ^2 = 70.9 P < 1%	F = 118.4 P = .0%
voc and gram	6	0	χ^2 = 1.8 P > 50%	F = 0 P = 84%
C-tests	5	0	χ^2 = 3.0 P > 50%	F = 0.1 P = 72%
dictation	7	5	χ^2 = 20.5 P < 1%	F = 150.6 P = 0%
translation	3	2	χ^2 = 6.8 P < 5%	F = 3.7 P = 6%
optimal comb	11	0	χ^2 = 7.4 P > 50%	F = 0 P = 93%

the optimal combination consists of the following tests:
voc (3) gra (3) C-Test (3) trans (2)

In the first line you see the results of our first global analysis of all our 21 subtests. 12 t-tests are significant, and the χ^2-Test and the TUKEY-Test are also significant. This means that the 21 subtests are <u>not</u> unidimensional, in other words, that there must be more than one language factor in this test battery.

Secondly, we analysed the six discrete-point tests measuring vocabulary and grammar with the CLA-model. You see that these tests are measuring the same thing.

The 5 C-Tests also constitute only one dimension.

But, as you see, the seven parts of the dictation are heterogenous; this test is measuring at least two different dimensions of language. Also the 3 translation tests are not unidimensional.

After this, we looked for an optimal combination of subtests. We made a number of attempts and at last we found an optimal combination.

If one adds the three discrete-point tests of grammar and the 3 tests of vocabulary, 3 of the 5 C-Tests and the 2 holistic trans-

BEST COPY AVAILABLE

lation tests, all these tests may be unidimensional (see Table 6). What about other dimensions?

I don't believe that the 2 remaining C-Tests constitute a second dimension (see line 3). The 7 dictation parts are heterogenous. Perhaps different parts of the original text measure different aspects of language. Here a further analysis would be useful.

Finally, the MC-translation, a new method of measuring translation competence, does not fit. Perhaps this subtest constitutes a further dimension. But if we look only at the 11 tests which fit the model, we can conclude:

the 3 discrete point grammar tests
the 3 discrete point vocabulary tests
the 3 holistic C-Tests
the 2 holistic translation tests

are assumed to be one-dimensional and interval scaled. The name of this factor could be: general language factor in written language.

Bibliography

Kalveram, K.T. (1966), "Die Veränderung von Faktorenstrukturen durch simultane Überlagerung". Arch. ges. Psychol. 117: 296-305.

Kalveram, K.T. (1970), "Problem der Selektion in der Faktorenanalyse". Arch. f. Psychol. 122: 199-230.

Moosbrugger, H. and Müller, H. (1982), "A Classical Latent Additive Test Model (CLA Model)". The German Journal of Psychology 6: 145-169.

Smith, V. and Klein-Braley, C. (1981), "Practical Approaches to the Teaching and Testing of Oral Skills for University Students in Germany". In: Klein-Braley, C. and Stevenson, D.K. (eds) (1981). Practice and problems in language testing. 1. Proceedings of the First International Language Testing Symposium of the IUS. Orbis Linguisticus 1. Frankfurt etc.: Verlag Peter D. Lang.

Smith, V. (1982), "On assessing speaking in small groups". In: Lutjeharms, M. and Culhane, T. (eds) (1982). Practice and problems in language testing 3. Proceedings of the Third International Language Testing Symposium of the IUS. Brüssel: Vrije Universiteit.

Thurstone, L.L. (1947). Multiple factor analysis. Chicago: University of Chicago Press.

BEST COPY AVAILABLE